DOI: 10.23977/erej.2025.090108
ISSN 2616-3756 Vol. 9 Num. 1

Analysis of influencing factors of air quality in Nanjing based on linear regression algorithm

Huanzheng Zhu, Jiaqiang Xie, Chenglong Chao, Zhengxun Fang

School of Mechanical and Electronic Engineering, Shandong Jianzhu University, Jinan, 250101, China

Keywords: Air Quality Index, Random Forest, Multiple Linear Regression, Grey Correlation

Abstract: With the advancement of urbanization and industrialization, energy consumption has increased, resulting in a surge in the discharge of harmful pollutants, and the problem of air pollution has become increasingly serious. The aim of this study is to provide decision support for improving air quality. By analyzing the data of AQI and six kinds of pollutants in Nanjing from 2018 to 2023, the current situation of air quality in Nanjing was explored, and the air quality level was evaluated from multiple perspectives. The factors affecting air quality were also discussed, and the correlation between pollutants, meteorological factors and economic factors and AQI was analyzed by using multiple linear regression, random forest and grey correlation analysis. The study found that the Air quality Index (AQI) in Nanjing was high in spring and winter and low in autumn and summer, and the air quality grade gradually improved from 2018 to 2022. According to the model regression, the key meteorological factors affecting the air quality of Nanjing are temperature and precipitation, and the important economic factors are the proportion of the secondary industry in GDP, the green coverage rate of built-up areas and the population density.

1. Introduction

Nowadays, the deteriorating air quality has seriously affected people's health and life, so how to improve air quality and promote urban ecological construction has become an urgent problem to be solved. As an important measure to deal with air quality problems in the era of science and technology, how to detect air quality has been paid more and more attention.

Cheng Hanxi et al. [1] analyzed the spatial and temporal distribution and correlation coefficient of AQI and air pollutants, and studied the change trend of air pollutant values and air quality in the Beijing-Tianjin-Hebei region in the past eight years. Huang Yu et al. [2] conducted multivariate statistical analysis based on the air quality data of Fuzhou for a total of 66 months from January 2014 to June 2019. Li Jiacheng et al. [3] selected AQI in Beijing from the first quarter of 2014 to the second quarter of 2022 as the research object to explore the impact of six major pollutants, five meteorological factors and 14 economic variables on air quality.

At present, most of the air quality research objects are relatively macro, such as the southwest region, the Yangtze River basin, the northeast region, etc., while Nanjing city is relatively lacking [4]. Therefore, this paper adopts Nanjing as the research object, which not only fills the relative blank of this block research, but also provides pertinence and reference value for the local area. In addition, most studies on influencing factors of air quality choose a single index for analysis, which is relatively one-sided, such as using the excellent and good rate of air quality grade or pollutants, while ignoring the role of other influencing factors. Therefore, by studying the impact of pollutants, meteorological factors and economic factors on air quality, this paper adopts machine learning algorithms such as multiple linear regression analysis, random forest and grey correlation analysis to analyze and grade air quality, and the results will be more scientific and reasonable.

2. Analysis of current situation of air quality in Nanjing

This article was collected through the China Air Quality Online Monitoring and Analysis platform (https://www.aqistudy.cn/). The collected data included the concentrations of AQI and six pollutants PM10, SO_2 and CO in Nanjing from 2018 to 2023.

2.1 Overview of air quality related indicators

Air Quality Index (AQI) can be used to evaluate air quality quantitatively, and the value of AQI reflects the degree of air pollution. The six detection parameters of the air quality standard are PM2.5, PM10, SO_2, NO_2, CO, O_3, and the six parameters all adopt a unified evaluation standard and are dimensionless values to reflect the air condition of the region. The AQI air quality levels in China is shown in Table 1.

Table 1 Air quality rating

Air Quality Level	AQI Range
Excellent	0-50
Good	51-100
Mild contamination	101-150
Middle level pollution	151-200
Serious contamination	201-300
Severe contamination	>300

The data used in this section are all from the Nanjing Air Quality online monitoring platform, covering a period of five years from 2018 to 2022. The variables included are AQI, date, six major pollutant concentrations and nine major indicators of quality grade.

2.2 Analysis of time change trend of AQI

2.2.1 Analysis of annual change trend of AQI

Based on the daily data of Nanjing AQI from 2018 to 2022, the trend chart of AQI was drawn to preliminarily judge the air quality of Nanjing. The change trend of Nanjing AQI in recent years is shown in Figure 1.

It can be seen from the annual change chart that the maximum change was in 2018, and the maximum AQI was also in that year. Nanjing's air quality has increased year by year, and the overall air quality in 2021 is relatively best in the study year. In the past five years, air quality was best in the middle of the year, and worse at the end and beginning of the year. Excluding extreme values, AQI curve has a certain regularity over the years. Compared to previous years, air quality has improved in 2022.

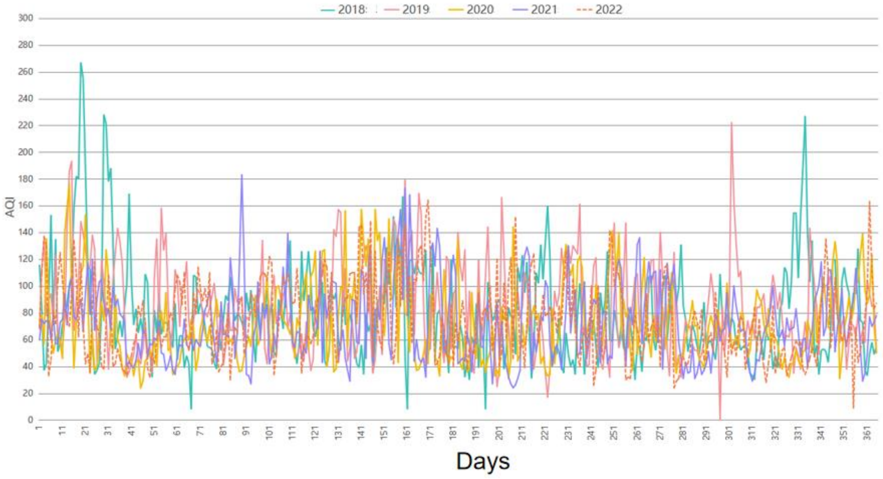


Figure 1 Changes in AQI from 2018 to 2022

2.2.2 Analysis of monthly and quarterly trends of AQI

According to the analysis of the AQI annual change curve above, the annual change of AQI presents a certain regularity. The monthly changes of AQI in Nanjing are shown in the Figure 2 below.

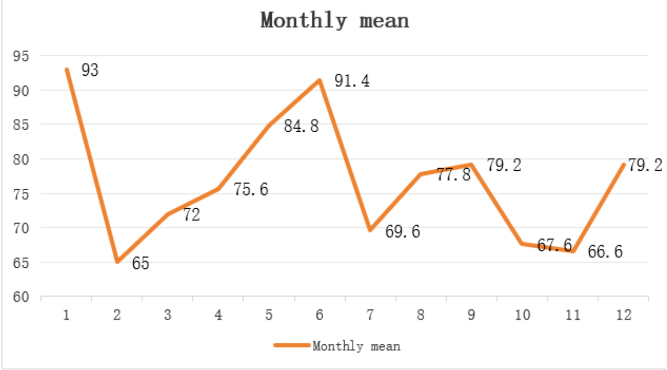


Figure 2. AQI monthly mean value diagram

The chart shows that the AQI monthly mean curve is roughly "W" shaped. The peak of AQI is in June, showing the characteristics of high in winter and summer and low in spring and autumn. AQI mutation decreased in January-February, then gradually rebounded and reached its peak in June. From July to December, the AQI value fluctuated with a small range.

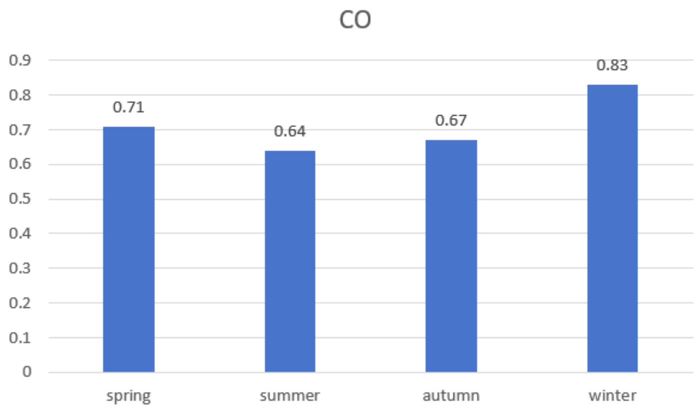


Figure 3. AQI quarterly mean value diagram

The month is planned as a quarter every four months, and the chart of AQI seasonal mean change is shown in Figure 3.

2.3 Distribution and analysis of air quality levels

The days with good air quality can evaluate the regional air quality and reflect the environmental quality and the health status of residents. Its monitoring helps the government and environmental protection departments to evaluate the effectiveness of environmental protection policies and take measures to improve air quality.

The following Figure 4 shows the specific distribution of air quality levels in Nanjing from 2018 to 2022.

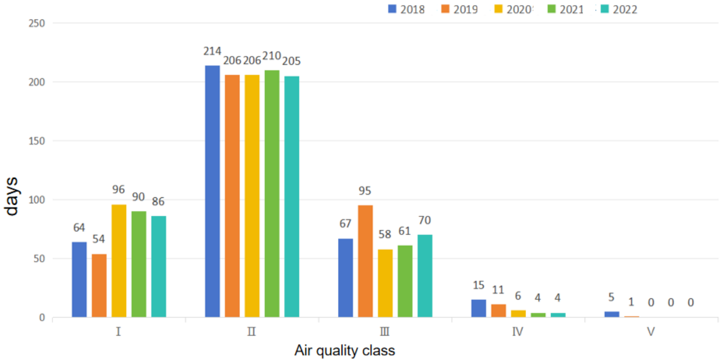
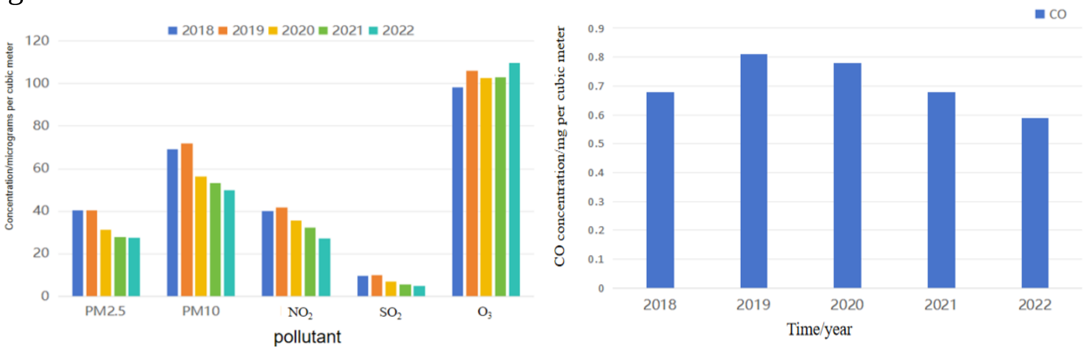


Figure 4 Distribution of air quality classes in Nanjing

As can be seen from the figure, the number of days with secondary air quality accounts for the highest proportion every year. After 2020, the number of high-grade air quality days increased significantly compared to previous years, indicating a decrease in air pollutants. From 2018 to 2022, the overall air quality level in Nanjing has gradually improved, indicating a better environment.

2.4 Analysis of the time change trend of the six major pollutants

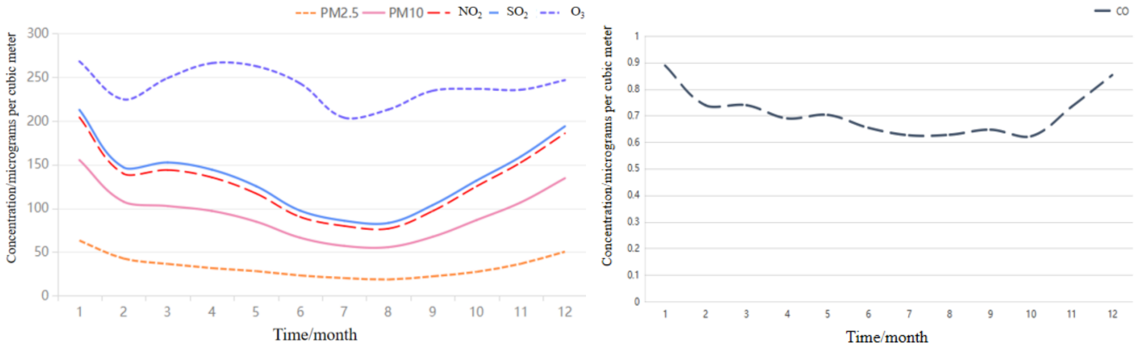
In order to study the concentration changes of the six major pollutants in Nanjing from 2018 to 2022, the annual and monthly mean values of the six major pollutants are plotted, as shown in Figure 5 and Figure 6.



(a) Column chart of pollutant change

(b) Column chart of CO changes

Figure 5. Annual mean changes of six major pollutants

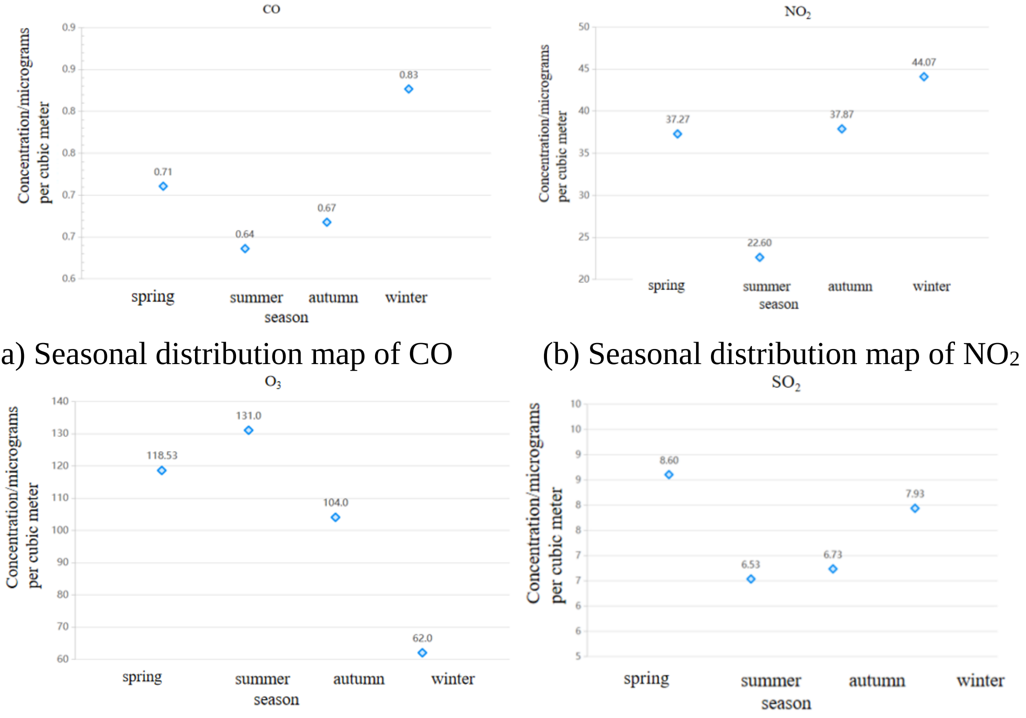


(a) Line chart of pollutant change (b) Line chart of CO change

Figure 6. Variation of monthly mean value of six major pollutants

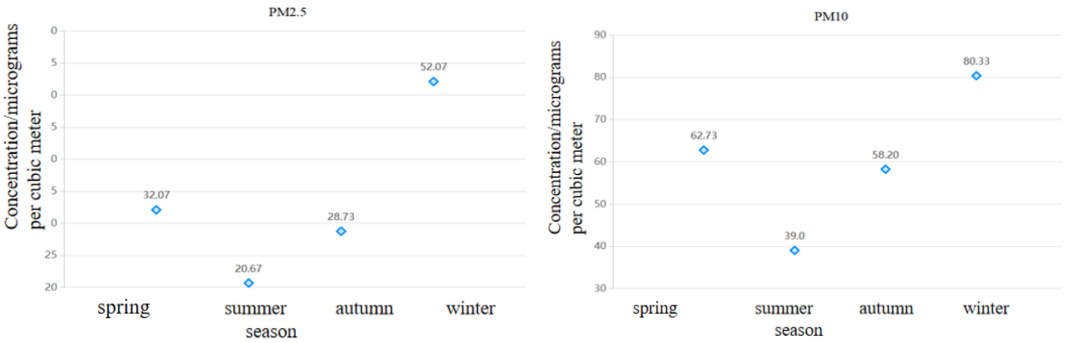
From the four figures, it can be seen that as a special pollutant, the concentration of CO is stable over the years and every month, almost in a straight line, indicating that its change is stable and difficult to control. The concentrations of PM2.5, PM10, SO_2 and NO_2 have decreased year by year, indicating that they have been effectively controlled. However, O_3 concentration is increasing year by year, and ozone is increasing overall, which reflects the destruction of ozone by human activities to a certain extent.

In order to specifically analyze the variation trends of the six pollutants in different seasons, the seasonal mean variation chart is drawn, as shown in Figure 7.



(a) Seasonal distribution map of CO (b) Seasonal distribution map of NO_2

(c) Seasonal distribution map of O_3 (d) Seasonal distribution map of SO_2



(e) Seasonal distribution of PM2.5 (f) Seasonal distribution of PM10

Figure 7. Variation of seasonal mean values of six major pollutants

From the six seasonal distribution maps, it can be seen that the seasonal concentrations of five pollutants, CO, NO_2, PM10, PM2.5 and SO_2, have similar trends, with high concentrations in spring and winter and low concentrations in autumn and summer. Ozone concentration is highest in summer, which is closely related to the use of chemicals such as freon to destroy the ozone layer.

In order to explore the primary pollutants in different seasons, grey correlation analysis was carried out. The analysis results are shown in the Table 2 below:

Table 2 Results of grey correlation analysis

Season	PM2.5	PM10	NO_2	CO	SO_2	O_3
Spring	0.885	0.921	0.90	1	0.674	0.697
Summer	0.406	0.42	0.417	0.688	0.653	0.566
autumn	0.838	0.899	0.680	0.991	0.964	0.814
Winter	0.348	0.484	0.579	0.692	0.922	0.397

From the table, we can intuitively understand the relationship between air quality and six major pollutants in different seasons. The correlation degree ranges from 0 to 1, and the higher the value, the closer the correlation. In spring, PM10, NO_2 and CO are the primary pollutants, which are highly correlated with each other. In autumn, the main pollutants were CO and SO_2, with a correlation of 0.99, almost highly correlated. In summer, CO and SO_2 were the main pollutants, but the correlation was low. In winter, the primary pollutant is SO_2, and it is highly correlated.

3. Results

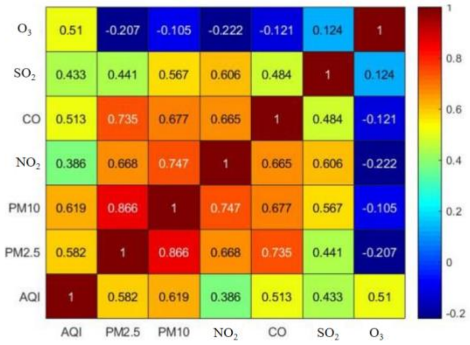
The status quo of air quality in Nanjing in the past five years has been analyzed, and the time change trend of air quality AQI and six major pollutants has been studied. This chapter continues to discuss the factors affecting Nanjing's air quality from three aspects of pollutants, meteorology and economy, and find out the main reasons.

3.1 The impact of pollutants on air quality

In this chapter, six most representative pollutants are selected as influencing factors, and their influence ratio on AQI is analyzed by multiple linear regression. The model can consider the effects of multiple independent variables on dependent variables at the same time, comprehensively analyze the multi-factor correlation and the interaction between independent variables, so as to better analyze the correlation of independent variables. By observing the correlation coefficient, we can understand the strength of correlation between independent variable and dependent variable. In this section, daily data from 2018 to 2022 are selected, and PM2.5, PM10, NO_2, CO, SO_2 and O_3 are recorded as independent variables X_1-X_6, and AQI is recorded as dependent variable Y, respectively. Multiple linear regression analysis is carried out using software.

3.1.1 Correlation analysis

In this section, Pearson correlation analysis is carried out for independent variables and dependent variables, and the thermal maps of correlation coefficients are shown in Figure 8.



	AQI	PM2.5	PM10	NO_2	CO	SO_2	O_3
O_3	0.51	-0.207	-0.105	-0.222	-0.121	0.124	1
SO_2	0.433	0.441	0.567	0.606	0.484	1	0.124
CO	0.513	0.735	0.677	0.665	1	0.484	-0.121
NO_2	0.386	0.668	0.747	1	0.665	0.606	-0.222
PM10	0.619	0.866	1	0.747	0.677	0.567	-0.105
PM2.5	0.582	1	0.866	0.668	0.735	0.441	-0.207
AQI	1	0.582	0.619	0.386	0.513	0.433	0.51

Figure 8. Heat map of correlation coefficient

As shown in the figure, the correlation coefficient between AQI and PM10 is 0.619, and that between AQI and PM2.5 is 0.582, showing a high correlation. The correlation coefficient with SO_2 was 0.433 and with NO_2 was 0.386, indicating a low correlation. The analysis shows that there is a correlation between AQI and the six major pollutants, which can be fitted by multiple linear regression.

3.1.2 The model is built on the test

Using Python software and Pandas and Scikit-learn library to build a multiple linear regression model, analyze the relationship between AQI (Air Quality Index) and six substances (PM2.5, PM10, NO_2, SO_2, CO, O_3), and print out the coefficient, intercept and relationship expression of the regression model. The output is as follows:

$$Y = 0.52X_1 + 0.27X_2 - 0.04X_3 - 0.36X_4 + 19.48X_5 + 0.44X_6 - 13.08 \tag{1}$$

In formula 1,

X_1 is PM2.5, X_2 is PM10, X_3 is N02, X_4 is SO_2, X_5 is CO,and X_6 is O_3.

According to the regression equation, PM2.5, PM10, CO and O_3 have a positive impact on AQI, and the regression coefficient of CO is the largest, which is 19.48, that is, when CO increases by one unit, AQI increases by 19.48 units on average. Therefore, improving air quality should focus on CO pollution. In this section, the three pollutants with the largest regression coefficients are selected: PM2.5, CO and O_3.

3.2 The impact of meteorological factors on air quality

In addition to pollutants, meteorological factors have a significant impact on air quality, which can directly affect air quality through temperature, humidity, wind speed, pressure and precipitation [4] .

In terms of temperature, usually the concentration of pollutants is positively correlated with the temperature. High temperature reduces the stable layer of the atmosphere, and the concentration of pollutants increases while it is difficult to diffuse. In terms of wind, high wind speed is conducive to the diffusion of pollutants, light wind and calm wind are generally beneficial, but strong winds may increase pollutants. Humidity is also key, low humidity, dust and other small affect breathing; High humidity limits the diffusion of pollutants. Precipitation can wash pollutants, reduce atmospheric concentration, but also conducive to plant growth, enhance the purification power of vegetation.

Meteorological factors play an important role and are subject to various conditions, which should be considered comprehensively. In this section, indicators such as average temperature, relative humidity, precipitation, and average wind speed are introduced. Then the random forest model was used to measure the importance of independent variables and rank the influencing factors of AQI. The model follows the random principle, can avoid multicollinearity, estimate the importance of features, and is suitable for many kinds of problems and data types, and is convenient for parallel computation. In this section, SPSS software is used to construct the model.

In order to verify the performance of the model, all the data were divided into the training set and the test set according to the ratio of 8:2, and the prediction error of the test set was obtained :MAE =0.938, MSE= 0.879, RMSE=0.938. The error was small, indicating that the model had good fitting performance and could be used to analyze the influencing factors of AQI.

Table.3. Results of model evaluation

	MSE	RMSE	MAE	MAPE
Training set	4.746	2.179	1.784	2.288
Test set	0.879	0.938	0.938	1.242

The following is the final data result:

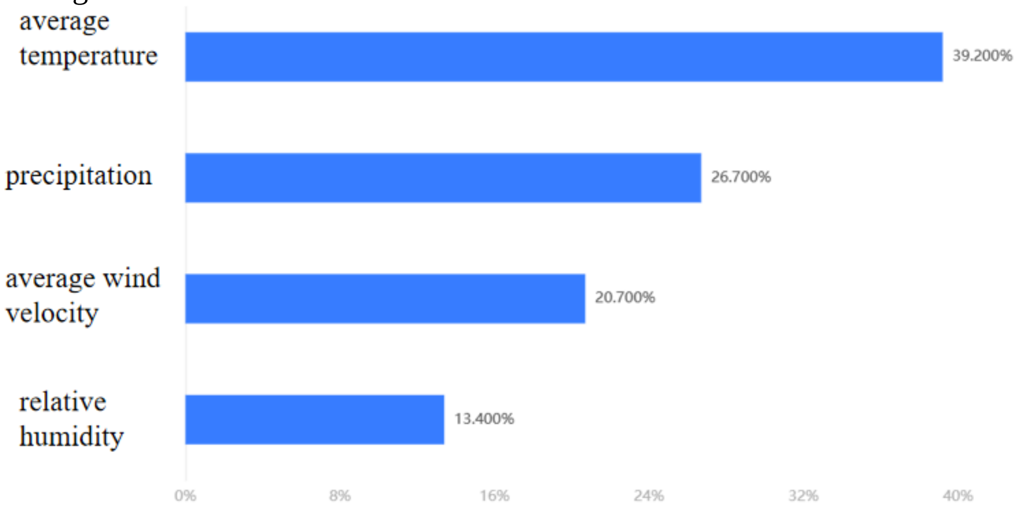


Figure 9. Ranking results of influence degree

Figure 9 or Table 3 shows the proportional importance of mean temperature, precipitation, mean wind speed, and relative humidity.

The final chart shows that among the importance scores of influencing factors, average temperature is the highest, followed by precipitation, then average wind speed, and relative humidity is the lowest. It can be seen that average temperature is the key factor affecting AQI, precipitation and average wind speed have significant effects, and relative humidity has a low impact.

3.3 The impact of economic factors on air quality

In addition to the natural environment, air quality is also affected by economic and social factors. Many experts have confirmed that the impact of socio-economic factors on air quality is complex and multi-dimensional, covering economic growth, industrial structure, greening, etc. [5] . However, it is relatively easy for relevant departments to control these social factors. Based on previous studies, this paper summarizes and analyzes six factors: per capital GDP, population density, green coverage rate of built-up areas, private car ownership, and electricity consumption of the whole society.

(1) Population density measures the population distribution of Nanjing. Although the high density increases the pollutants due to the energy consumption of economic development, it also reflects the advantages of regional development, or is conducive to the improvement of air quality. (2) Per capital GDP reflects the economic and income level of Nanjing. Economic development increases pollution, while the enhancement of environmental awareness helps alleviate it. (3) The green coverage rate of built-up areas measures urban greening, and improving it is beneficial to air quality, climate and environment. (4) The number of private cars reflects the development of transportation, and the increase of private cars leads to the increase of pollutant emissions, damaging the atmosphere and reducing air quality. (5) The proportion of GDP of the secondary industry reflects the industrial structure and industrialization degree, and its energy consumption and construction dust are the main causes of air pollution. (6) Nanjing mainly uses photovoltaic, wind power and other green power generation, which is beneficial to the air; In some areas, traditional thermal power consumption is high and pollution is heavy.

The data used this time are from Nanjing Statistical Yearbook. SPSS software was used to analyze the social factors affecting the air quality in Nanjing. Every two data were taken as a group, and the personal correlation coefficient was selected to calculate the correlation degree. The test results are shown in the following Table 4.

Table 4 Phase relation table

	AQI	Total electricity consumption	Civilian private car ownership	The proportion of secondary industry in GDP	Per capital GDP	Population density	Green coverage rate of built-up area
AQI	1***						
Total electricity consumption	-0.557	***					
Civilian private car ownership	-0.598	0.99***	1***				
The proportion of secondary industry in GDP	0.426	-0.221	-0.335	1***			
Per capital GDP	-0.319	0.925**	0.898**	-0.244	1***		
Population density	0.07	-0.63	-0.64	-0.075	-0.469	1***	
Green coverage rate of built-up area	-0.272	0.404	0.473	-0.893**	0.559	0.055	1***
Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively							

According to the correlation table, the population density and the proportion of secondary industry GDP in Nanjing are positively and linearly correlated with AQI, and the correlation coefficients are 0.07 and 0.426 respectively, indicating that the correlation of population density is small, and the secondary industry GDP has a great impact on air quality. The larger the proportion, the lower the air quality is generally. Private car ownership and electricity consumption in the whole society are highly negatively correlated with AQI, indicating that the improvement of living standards will lead to air pollution. However, the personal coefficient only reflects the relationship between pairwise factors and does not reflect the influence of multi-factor synthesis on AQI, while the influence of social and economic factors on AQI is complex, and only the correlation coefficient analysis is insufficient. Therefore, grey correlation analysis is used to calculate the correlation degree between each factor and AQI to find out the key factors affecting air quality.

Gray correlation analysis is not limited by data distribution, suitable for linear and nonlinear relationships, can process all kinds of data, including quantitative and qualitative data, has a certain resistance to noise and interference, and can effectively deal with incomplete or inaccurate data. The analysis results are shown in Table 5 below.

Table 5 Results of grey correlation analysis

Correlation result		
Evaluation item	Degree of association	Ranking
The proportion of secondary industry in GDP	0.761	1
Green coverage rate of built-up area	0.662	2
Population density	0.578	3
Per capital GDP	0.577	4
Civilian private car ownership	0.556	5
Total electricity consumption	0.532	6

The results of grey correlation degree analysis show that the correlation degrees from high to low are: secondary industry, green coverage rate of built-up areas, population density, per capita GDP, private car ownership, and electricity consumption of the whole society. Among them, the correlation between the secondary industry and the green coverage rate of built-up areas is quite high, reaching 0.761 and 0.662 respectively. The secondary industry covers electricity, industrial manufacturing, gas and other sectors, and its development is prone to increase pollutants. The correlation degree of the other four factors is below 0.6, which belongs to moderate and mild correlation, and the influence on AQI is not significant, with certain deviation.

4. Conclusions

Based on the data of Nanjing from 2018 to 2022, this paper analyzes the current situation of air quality and key influencing factors, and uses random forest and multiple linear regression to draw conclusions and suggestions: Through descriptive analysis, AQI in Nanjing is high in spring and winter, low in autumn and summer, and the primary pollutant in winter is SO_2. From 2018 to 2022, air quality improved and the concentration of four pollutants, including PM2.5, decreased. The model regression points out that the key meteorological factors are temperature and precipitation, and the economic factors are the proportion of the GDP of the secondary industry, the green coverage rate of the built-up area and the population density. In terms of pollutants, the government should focus on controlling PM2.5, CO and O_3. Reduce coal burning at source, promote clean heating, control volatile organic compounds, control high-carbon emission industries and industrial emissions, strengthen exhaust control, adjust the energy mix, and formulate regional policies. Meteorological factors are difficult to control. Based on economic factors, to improve air quality, it is necessary to increase the proportion of GDP of the secondary industry, the green coverage rate of built-up areas and reasonable population control. The government can introduce industrial policies, strengthen scientific and technological innovation, promote environmental protection and resource conservation, increase investment in greening, and protect ecological resources.

In the actual air quality detection, it will be affected by many factors. Therefore, the main influencing factors should be selected for detection according to the actual local situation. Relevant departments should comprehensively introduce air quality factors. This is conducive to obtaining highly accurate detection data and making reasonable and scientific decisions.

References

[1] Cheng Hanxi, Zhu Hongxia, Wang Jing, et al. Impact of air pollution control on air quality in Beijing-Tianjin-Hebei region [J]. Environmental Impact Assessment, 2019, 46(06):78-85.

[2] Huang Yu, Liu Jinfu, You Tiange, et al. Air quality detection in Fujian Province based on multivariate statistics [J]. Regional Governance, 2019, (42):45-47.

[3] Li Jiacheng, Liang Longyue. Air quality prediction and Influencing factor identification based on Machine learning method [J]. Computer Technology and Development, 2019, 34(01):164-170.

[4] Ren J H. Evaluation method and application of meteorological factors on atmospheric pollutant concentration [D]. Chinese Academy of Environmental Sciences, 2024.

[5] Huang Qiaolong, Cai Xuexiong. Digital economy development and air quality improvement: An analysis based on innovation-driven perspective [J]. Journal of Enterprise Economics, 2018, 43(07):91-101.